УДК 519.28

ПОВЫШЕНИЕ БЫСТРОДЕЙСТВИЯ ТОЧНОГО АЛГОРИТМА РЕАЛИЗАЦИИ МЕТОДА НАИМЕНЬШИХ МОДУЛЕЙ ПРИ ОЦЕНИВАНИИ ПАРАМЕТРОВ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

ENHANCING THE PERFORMANCE OF THE EXACT ALGORITHM FOR IMPLEMENTING THE METHOD OF LEAST ABSOLUTE DEVIATIONS WHEN ESTIMATING THE PARAMETERS OF LINEAR REGRESSION MODELS

Азарян Алексан Артурович, аспирант каф. «Прикладная математика», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: a.a.azaryan@gmail.com, Тел.: +7(902)257-32-33
Тырсин Александр Николаевич, д-р. техн. наук, заведующий каф. «Прикладная математика», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: at2001@yandex.ru. Тел.: +7(922)100-74-52

Aleksan A. Azaryan, PhD student, Department « Applied mathematics », Ural Federal University named after the first President of Russia B.N.Yeltsin, 620002, Mira street, 19, Ekaterinburg, Russia. E-mail: a.a.azaryan@gmail.com. Ph.: +7(902)257-32-33
Alexander N. Tyrsin, Doctor Sc., Department Chairman «Applied mathematics», Ural Federal University named after the first President of Russia B.N.Yeltsin,620002, Mira str., 19, Ekaterinburg, Russia. E-mail: at2001@yandex.ru. Ph.: +7(922)100-74-52

Аннотация: Описан алгоритм нахождения точного решения задачи оценивания параметров линейных регрессионных моделей методом наименьших модулей. Данный алгоритм значительно выигрывает по сравнению с известным переборным алгоритмом. Приведены сравнительные характеристики предложенного и известного алгоритмов. Описан пример практической реализации алгоритма.
Abstract: An algorithm for finding the exact solution of the problem of estimating the parameters of linear regression models by the method of least absolute deviations is described. This algorithm significantly wins in comparison with the known search algorithm. Comparative characteristics of the proposed and known algorithms are given. An example of practical implementation of the algorithm is described.
***Ключевые слова*:** метод наименьших модулей; линейная регрессионная модель; алгоритм; оценка.
***Key words*:** the method of least absolute deviations; linear regression model; algorithm; estimation.

ВВЕДЕНИЕ

Одной из типовых задач при статистической обработке результатов экспериментальных исследований является задача оценивания коэффициентов линейной регрессионной модели

$$y = Xa + \varepsilon, \qquad (1)$$

где y— вектор наблюдаемых значений зависимой переменной; $X = \{x_{ij}\}_{n\times m}$— матрица значений объясняющих (независимых) переменных; ε — вектор случайных погрешностей (ошибок) измерений; a — вектор коэффициентов регрессии.

Среди прикладных статистических методов построения регрессионных зависимостей наибольше распространение получил метод наименьших квадратов (МНК). Использование МНК требует выполнения ряда предпосылок, называемых условиями Гаусса-Маркова [1, 2]. При их выполнении МНК-оценки параметров модели (1) являются состоятельными и несмещенными. Кроме того, если случайные ошибки имеют нормальный закон распределения, то МНК-оценки становятся эффективными. Однако использование МНК при нарушении условий Гаусса-Маркова может привести к значительным ошибкам при оценивании коэффициентов [3, 4].

МЕТОДИКА РЕШЕНИЯ

С целью обеспечения устойчивости оценок относительно отклонений случайных ошибок от

гауссовой модели разработан ряд статистических методов. Данные методы основаны на более общих предположениях относительно случайных ошибок. Одним из таких методов является метод наименьших модулей (МНМ) [5]. МНМ для задачи (1) имеет вид

$$a^* = \arg\min_{a_1,\cdots,a_m} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} a_j x_{ij} \right|, \quad (2)$$

где $a^* = (a_1^*, a_2^*, \cdots, a_m^*)$ — вектор оценок параметров модели (1).

Известный точный метод решения задачи (2), основанный на переборе всех узловых точек [6] требует решения C_n^m систем линейных уравнений порядка m. Вычислительные погрешности здесь не значительны. Однако с ростом n и m наблюдается экспоненциальный рост вычислительных затрат. Использование современных вычислительных технологий, например параллельного программирования, лишь частично решает указанную проблему за счет использования нескольких процессоров, приводя к значительному усложнению реализации и удорожанию вычислительного процесса.

Рассмотрим алгоритм точного решения задачи (2), существенно выигрывающий по объему вычислительных затрат по сравнению с переборным алгоритмом.

Алгоритм основан на спуске к точному решению, двигаясь вдоль прямых l, каждая из которых является пересечением $(m-1)$ различных гиперплоскостей p_i, которые построены по данным из выборки:

$$p_i : y_i - \sum_{j=1}^{m} a_j x_{ij} = 0,$$

$$l_{(k_1,\cdots,k_{m-1})} : \bigcap_{i=k_1}^{k_{m-1}} p_i \ , k_j \in \{1,2,\cdots,n\}.$$

В качестве начального приближения берется произвольная узловая точка, являющаяся пересечением m гиперплоскостей $p_{k_1},\cdots,p_{k_m}$. Исключив одну из гиперплоскостей, получим прямую l. В любой узловой точке можно построить m таких узловых прямых. Выберем ту, вдоль которой целевая функция достигает наименьшего значения, которое всегда будет достигаться в одной из узловых точек. Найдя эту точку, продолжим движение из нее по тому же принципу. В результате будет найдена узловая точка, спуск из которой невозможен. И эта узловая точка будет являться точным решением задачи (2). Существование решения вытекает из выпуклости целевой функции $Q(a) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} a_j x_{ij} \right|$.

Поясним этот алгоритм, используя рис. 1. В качестве начальной точки выберем точку *H*. Через нее проходят прямые I и II. Сначала рассмотрим прямую I. Среди узловых точек, которые лежат на этой прямой, выбираем ту, в которой достигается минимальное значение целевой функции (таким образом, происходит спуск по прямой). Предположим, что такой точкой является точка *C*. Рассмотрим прямую II. Спускаясь по прямой II, находим на ней точку, в которой достигается минимальное значение. Пусть, например, эта точка *L*. Далее, сравнивая значения целевой функции в точках *C* и *L*, выбираем ту, в которой достигается минимальное значение. Пусть, например, эта точка *C*. Через нее помимо прямой I проходит прямая III. На этой прямой находим очередную точку, в которой достигается минимальное значение целевой функции. Пусть такой точкой является *D*. Допустим, что на второй прямой, проходящей через *D*, точкой минимума является сама точка *D*, тогда в качестве решения выбираем точку *D*, в которой достигается минимум целевой функции (если бы нашлась другая точка минимума на этой прямой, то из нее продолжили бы дальнейший спуск).

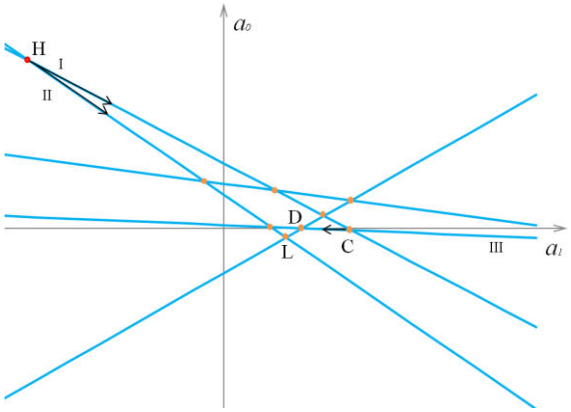

Рис 1. Спуск по узловым прямым

Двигаясь вдоль прямых l, для нахождения узловых точек, которые лежат на этой прямой, мы каждый раз решаем систему линейных уравнений (СЛУ) порядка m. Очевидно, что СЛУ двух различных узловых точек, которые лежат на одной прямой ($l_{(k_1,\cdots,k_{m-1})}$), отличаются лишь одним уравнением. Действительно они имеют вид

$$\begin{cases} p_{(k_1)} = 0 \\ p_{(k_2)} = 0 \\ \cdots \\ p_{(k_{m-1})} = 0 \\ p_{(k_i)} = 0 \end{cases} \text{ и } \begin{cases} p_{(k_1)} = 0 \\ p_{(k_2)} = 0 \\ \cdots \\ p_{(k_{m-1})} = 0 \\ p_{(k_j)} = 0 \end{cases},$$

где $k_i, k_j \notin \{k_1, k_2, \cdots, k_{m-1}\}$ и $k_i \neq k_j$.

Следовательно, вычислительная эффективность алгоритма спуска существенно повысится, если для нахождения узловых точек, которые лежат на

прямой $l_{(k_1,\cdots,k_{m-1})}$ мы найдем разреженную матрицу СЛУ этой прямой и, используя ее, решим соответствующее уравнение $p_{(k_i)} = 0$, где $k_i \notin \{k_1, k_2,\cdots, k_{m-1}\}$.

Расширенная матрица СЛУ прямой $l_{(k_1,\cdots,k_{m-1})}$ имеет вид

$$A_{l_{(k_1,\cdots,k_{m-1})}} = \begin{pmatrix} x_{k_1 1} & \dots & x_{k_1 m} & y_{k_1} \\ x_{k_2 1} & \dots & x_{k_2 m} & y_{k_2} \\ \dots & \dots & \dots & \dots \\ x_{k_{m-1} 1} & \dots & x_{k_{m-1} m} & y_{k_{m-1}} \end{pmatrix}.$$

Применив алгоритм прямого хода метода Гаусса, вместо исходной расширенной матрицы получится трапециевидная:

$$A'_{l_{(k_1,\cdots,k_{m-1})}} = \begin{pmatrix} 1 & x'_{k_1 2} & \dots & x'_{k_1 m} & y'_{k_1} \\ 0 & 1 & \dots & x'_{k_2 m} & y'_{k_2} \\ \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots 1 & x'_{k_{m-1} m} & y'_{k_{m-1}} \end{pmatrix}.$$

Используя разреженную матрицу $A'_{l_{(k_1,\cdots,k_{m-1})}}$ мы быстрее находим все узловые точки, которые лежат на прямой $l_{(k_1,\cdots,k_{m-1})}$.

Вычислительная эффективность алгоритма спуска можно повысить, рассматривая направление спуска. Поясним как.

Упорядочиваем все узловые точки, которые лежат на одной прямой, и выполняем вышеописанный алгоритм спуска, но с учетом направления. Если при непосредственном переходе от одной узловой точки к другой значение целевой функции увеличивается, то в этом направлении значение целевой функции будет увеличиваться во всех узловых точках (вытекает из выпуклости целевой функции). Назовем такое направление "плохим". Для осуществления спуска, до вычисления значения целевой функции в очередной узловой точке рассматриваем направление спуска. Если оно "плохое", то переходим к следующей точке без вычисления значения целевой функции в данной узловой точке.

Таблица 1.

Продолжительность работы различных алгоритмов (в сек.) при m=5 - количество переменных и N=100 - число испытаний

n - объем выборки	Переборный алгоритм	Алгоритмы спуска по узловым точкам		
		Обычный спуск	Спуск с использованием разреженных матриц	Спуск с использованием разреженных матриц и с учетом направления спуска
30	625	5.6	4.5	3.94
100	353418	34.14	24	20.5
300	92011252	152.74	101	78.7

В программной среде R была разработана программа, которая реализует переборный алгоритм и алгоритмы спуска по узловым точкам. Данные о продолжительности работы данных алгоритмов приведены в таблице 1. Эти данные показывают, что алгоритмы спуска существенно выигрывают по объему вычислительных затрат по сравнению с переборным алгоритмом.

ЗАКЛЮЧЕНИЕ

Предложен эффективный алгоритм реализации метода наименьших модулей для оценивания параметров многомерных линейных регрессионных моделей. Он представляет собой спуск по узловым точкам. Причем направление спуска задается по узловым прямым. Нахождение точного решения имеет вычислительные затраты, сравнимые с методом покоординатного спуска. Можно отметить значительное снижение вычислительных затрат при реализации обобщенного метода наименьших модулей. Быстродействие алгоритма спуска по узловым точкам может быть в дальнейшем увеличено.

Исследование выполнено при поддержке РФФИ, грант № 16-06-00048а.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Демиденко Е.З. Линейная и нелинейная регрессия. М.: Финансы и статистика, 1981. 302 с.
2. Кибзун А.И., Горяинова Е.Р., Наумов А.В. Теория вероятностей и математическая статистика. Базовый курс с примерами. 2-е изд. М.: ФИЗМАТЛИТ, 2005. 232 с.
3. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. М.: Финансы и статистика, 1987. 239 с.
4. Смоляк С.А. Устойчивые методы оценивания. М.: Статистика, 1980. 208 с.
5. Мудров В.И., Кушко В.Л. Методы обработки измерений. Квазиправдоподобные оценки. М.: Радио и связь, 1983. 304 с.
6. Тырсин А.Н. Робастное построение регрессионных зависимостей на основе обобщенного метода наименьших модулей // Записки научных семинаров ПОМИ. 2005. Т. 328. С. 236–250.